文章编号:1000-0887(2019)11-1179-25

近似 Bayes 计算前沿研究进展及应用*

朱万闯[1], 季春霖[2], 邓 柯[1]

(1. 清华大学 工业工程系 统计学研究中心, 北京 100084;
2. 光启高等理工研究院, 广东 深圳 518000)

摘要: 在大数据和人工智能时代,建立能够有效处理复杂数据的模型和算法,以从数据中获取有用的信息和知识是应用数学、统计学和计算机科学面临的共同难题.为复杂数据建立生成模型并依据这些模型进行分析和推断是解决上述难题的一种有效手段.从一种宏观的视角来看,无论是应用数学中常用的微分方程和动力系统,或是统计学中表现为概率分布的统计模型,还是机器学习领域兴起的生成对抗网络和变分自编码器,都可以看作是一种广义的生成模型.随着所处理的数据规模越来越大,结构越来越复杂,在实际问题中所需要的生成模型也变得也越来越复杂,对这些生成模型的数学结构进行精确地解析刻画变得越来越困难.如何对没有精确解析形式(或其解析形式的精确计算非常困难)的生成模型进行有效的分析和推断,逐渐成为一个十分重要的问题.起源于 Bayes 统计推断,近似 Bayes 计算是一种可以免于计算似然函数的统计推断技术,近年来在复杂统计模型和生成模型的分析和推断中发挥了重要作用.该文从经典的近似 Bayes 计算方法出发,对近似 Bayes 计算方法的前沿研究进展进行了系统的综述,并对近似 Bayes 计算方法在复杂数据处理中的应用前景及其和前沿人工智能方法的深刻联系进行了分析和讨论.

关 键 词: 近似 Bayes 计算; 生成模型; 深度学习; 不确定性推断

中图分类号: O357.41 **文献标志码:** A DOI: 10.21656/1000-0887.400245

引 言

在大数据和人工智能时代,建立能够有效处理复杂数据的模型和算法,以从数据中获取有用的信息和知识是应用数学、统计学和计算机科学面临的共同难题.在数据建模和分析的诸多技术路线中,基于"生成模型"(generative models)的数据分析方法是十分重要的一个研究领域.一般而言,生成模型是指用定量方式来描述数据的生成机制,其核心特征和基本目标是能够从给定的模型出发模拟产生和所研究的实际数据相似的模拟数据.由生成模型模拟产生的数据一般具有多样性,会形成一个数据空间上的概率分布 $F_{\boldsymbol{\theta}}$,其具体形式由模型参数 $\boldsymbol{\theta}$ 所决定.

在不同的学科领域,生成模型呈现出不同的表现形态.在应用数学研究中,数学家常常运用微分方程(differential equations)和动力系统(dynamic systems)等数学工具对产生数据的物

* **收稿日期:** 2019-08-26; **修订日期:** 2019-09-01
基金项目: 国家自然科学基金(11771242)
作者简介: 朱万闯(1988—),男,博士(E-mail: wanchuang.zhu@ sydney.edu.au);
季春霖(1981—),男,正高级工程师,博士(E-mail: chunlin.ji@ kuang-chi.org);
邓柯(1982—),男,副教授,博士,博士生导师(通讯作者. E-mail: kdeng@ tsinghua.edu.cn).

理过程进行精细描述[1-5];在统计学研究中,统计学家借助概率分布建立统计模型(statistical models)对观测数据的生成机制进行近似刻画[6-8];在机器学习研究中,人们运用生成对抗网络(generative adversarial networks, GAN)[9]、变分自编码器(variational autoencoders, VAE)[10]等技术手段,通过数据驱动的方式对图像、音频和文本等多种形式数据的产生过程加以人工模拟[11-21].从宏观的角度看,上述这些精细描述、近似刻画、人工模拟数据产生过程的不同方法,都是生成模型的某种具体形式.

上述不同形态的生成模型在建模原则、所用信息、可解释性、稳定性、适用场景等方面有不同的特性.以微分方程和动力系统为代表的"精细模型"(fine models)从微观入手对产生数据的物理过程给予定量刻画,力求细致准确;建模所需的信息以领域知识为主,需要对实际的物理机制有很深的了解;其可解释性较好,但一般仅适用于物理机制清晰明确的工程问题,应用于带有复杂结构的问题时往往面临较大的困难,且比较容易受到噪声的干扰.统计模型则在保留数据核心特征的同时对次要特征和噪声进行简化,以在模型的精准性和复杂性之间达到适当的平衡;通过将领域知识和数据特征相结合的方式来进行建模,可解释性和稳定性较好;一般适用于数据生成机制和噪声模式比较清晰的科学问题,特别是通过实验设计和抽样调查等规范手段收集数据的问题;但需要具体问题具体分析,建模成本较高.以 GAN 和 VAE 为代表的"数据驱动生成模型"(data-driven generative models)主要通过数据驱动的方式,从一个很大的模型空间中优化选取适当的模型来完成建模;对领域知识要求较低,适应性强,在图像、音频和文本等数据生成机制用常规方法难以刻画的问题中取得了良好的实际效果;但可解释性和稳定性相对较差.

尽管上述这些生成模型有着多样化的形态和差异化的特性,但是它们都有一个共同的作用:连接数据和科学问题的纽带,将数据转化为有用知识的桥梁.无论是以哪种形态出现的生成模型,其模型参数都在不同程度上或直接或间接地反映了所关心的科学问题的关键信息和核心知识.以生成模型为基础的数据分析,其本质是依据已知的观测数据对未知的模型参数进行推断,从而将数据转化为知识并对所关心的科学问题给予解答.在应用数学和机器学习中,人们常常运用最优化的手段对未知的模型参数进行估计,即在参数空间中找到满足特定最优化准则的最优解来作为未知参数的估计值.在统计学研究中,人们除了关心"最优估计",还特别关注对未知参数的不确定性进行度量和推断.从科学哲学的角度来看,不确定性度量和推断是一切数据分析都不可避免的一个根本问题.原因在于,尽管"最优解"从某种程度上可以对观测数据给出"最佳"的解释,但是其他非最优解也同样可以对观测数据给出较为合理的解释,因而并不能被决然排除,从而必然给未知参数的推断带来"不确定性".因而,在一切基于生成模型的数据分析中,不确定性推断都是一个十分重要、不容回避的问题.

相对于单纯的优化,不确定性推断在哲学思辨上要困难很多,技术上也更为复杂.在统计学框架下,有不同的范式来进行不确定性推断,较为常见的是频率学派(frequentist)和 Bayes 学派(Bayesian).Bayes 学派对未知参数设置先验分布(prior distribution),并通过 Bayes 公式(Bayes formula)对数据似然(data likelihood)实现反转而生成后验分布(posterior distribution).作为参数空间上的一个概率分布,后验分布对未知参数的不确定性进行了完整的刻画;作为给定观测数据的一个条件概率,后验分布充分整合了先验分布和观测数据中所包含的关于未知参数的信息.因而,Bayes 框架具有逻辑清晰简明、操作方便快捷的优势,在众多数据分析问题中得到了广泛的应用.然而,经典 Bayes 分析要求数据似然必须具有明确的解析形式.这一要求一方面给 Bayes 分析赋予了严格的概率含义,但同时也限制了 Bayes 分析在许多问题中的应

用:在许多复杂的实际问题中,数据似然难以被精确刻画和计算的情况时有发生.本文的第 1.3 小节将对此类现象给出详细的说明.同时,由于复杂问题中的后验分布大多不再是常规的标准分布,需要通过分布近似或者 Monte-Carlo 抽样的方式来进行处理和分析,Bayes 方法在计算上往往面临较大的挑战.

为了简化 Bayes 计算,特别是在数据似然难以被精确刻画和计算的情况下运用 Bayes 框架来进行数据分析,近似 Bayes 计算(approximate Bayesian computation, ABC)方法进入了人们的视野.ABC 方法的思想最初由 Rubin 在 1984 年提出[22],随后在许多领域得到了成功的应用,包括群体遗传学(population genetics)研究[23]、考古学(archaeology)[24]、金融建模(financial modelling)[25]、水文模型(hydrological models)[26]、蛋白质网络(protein networks)[27]等.相对于经典的 Bayes 推断和计算框架,ABC 方法的一个重要贡献是在保持了 Bayes 分析基本框架和概率解释的同时解除了对数据似然精确解析形式的强制依赖.通过从数据中提取特定的统计量,并在统计量之间定义适当的距离来刻画一组新生成数据和观测数据之间的相似性,ABC 方法可以在数据似然解析形式不明确的情况下对真实后验分布给出有效近似.ABC 方法的这一特性可以将 Bayes 分析框架从经典的统计模型推广到一般的生成模型,为 Bayes 分析和人工智能方法的结合提供新的机遇[28-31].

本文将系统介绍和综述 ABC 方法经典框架和前沿进展,并深入讨论 ABC 方法与前沿人工智能技术有机结合的思想、方法和应用.本文的剩余部分将按以下的方式安排:第 1 节简要介绍经典 ABC 方法的研究背景、基本框架和理论性质;第 2 节从多个角度综述近年来 ABC 方法研究的前沿进展;第 3 节讨论了 ABC 方法在复杂数据处理,特别是人工智能中的潜在应用;第 4 节对本文进行总结.

1 经典 ABC 方法简介

1.1 Bayes 推断的基本范式

令 $\boldsymbol{y}_n=(y_1,y_2,\cdots,y_n)$ 代表观测数据集,$\boldsymbol{\theta}\in R^p$ 表示要学习的参数,p 是参数的维度.现实中,观测数据集的生成方式可能十分复杂.统计模型通过对数据的生成方式添加理想化的假设条件达到简化数据生成方式的目的.通常,我们假设 $\boldsymbol{y}_n$ 是随机变量 Y 的一组实现.并且,随机变量 Y 服从与 $\boldsymbol{\theta}$ 相关的分布:$Y\sim F(\boldsymbol{\theta})$,其中分布的概率密度函数由 $f(\cdot\mid\boldsymbol{\theta})$ 表示.统计推断是根据观测数据集来推断参数 $\boldsymbol{\theta}$ 的大小.举个例子,假设 $\boldsymbol{y}_n$ 是通过简单随机抽样的方式,在全国范围内收集 n 个成年人的身高数据.我们的目标是推断全国成年人的平均身高 $\theta\in\mathbf{R}$.在生物学上,人类的身高可能受到多方面的影响,比如,基因差异、营养水平、气候等.但是,统计模型把身高假设为一个服从正态分布的随机变量.即身高 $Y\sim N(\theta,\sigma^2)$,其中 σ^2 是已知量.通过上述假设,统计模型将现实世界的各种因素都包含在统计分布中.

给定上述的统计模型,观测数据集的似然函数可以写为 $L(\boldsymbol{\theta})=f(\boldsymbol{y}_n\mid\boldsymbol{\theta})$.频率学派认为 $\boldsymbol{\theta}$ 是一个未知但是固定值;而 Bayes 学派则把 $\boldsymbol{\theta}$ 当做是一个随机变量,并用一个“先验分布” $\pi_0(\boldsymbol{\theta})$ 来表示在观测到数据之前我们对参数 $\boldsymbol{\theta}$ 的先验知识.Bayes 推断是基于如下的 Bayes 公式:

$$\pi_n(\boldsymbol{\theta})=f(\boldsymbol{\theta}\mid\boldsymbol{y}_n)=\frac{f(\boldsymbol{y}_n\mid\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})}{\int f(\boldsymbol{y}_n\mid\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}},\tag{1}$$

其中,$\pi_n(\boldsymbol{\theta})$ 被称为未知参数 $\boldsymbol{\theta}$ 的“后验分布”,即在观测到数据 $\boldsymbol{y}_n$ 后对未知参数 $\boldsymbol{\theta}$ 的认识.作

为参数空间上的一个概率分布,后验分布对未知参数的不确定性进行了完整的刻画;作为给定观测数据的一个条件概率,后验分布充分整合了先验分布和观测数据中所包含的关于未知参数的信息.因而,Bayes 框架具有逻辑清晰简明、操作方便快捷的优势,在众多数据分析问题中得到了广泛的应用.

如下定理为 Bayes 推断给出了基本的理论保证.

定理 1 令 $\boldsymbol{\theta}_0$ 代表参数的真实值,$\hat{\boldsymbol{\theta}}_n$ 代表似然方程的强相合解(strongly consistent solution).定义 $\boldsymbol{t}=\sqrt{n}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_n)$,并令 $\pi_n(\boldsymbol{t}\mid \boldsymbol{y}_n)$ 为 $\boldsymbol{t}$ 的后验分布.当正则条件(A1)~(A5)成立时,下式以概率 1 成立:

$$\lim_{n\to\infty}\int_{R^p}\left|\pi_n(\boldsymbol{t}\mid \boldsymbol{y}_n)-\frac{\sqrt{\boldsymbol{I}(\boldsymbol{\theta}_0)}}{\sqrt{2\pi}}\mathrm{e}^{-t^2\boldsymbol{I}(\boldsymbol{\theta}_0)/2}\right|\mathrm{d}\boldsymbol{t}=0, \tag{2}$$

其中 $\boldsymbol{I}(\boldsymbol{\theta})=E_{\boldsymbol{\theta}}\left(-\dfrac{\partial^2}{\partial\theta_i\partial\theta_j}\ln f(\boldsymbol{y}_n\mid\boldsymbol{\theta})\right)$.

在参数 $\boldsymbol{\theta}$ 的维度 $p=1$ 的情况下,正则性条件(A1)~(A5)的具体形式如下:

(A1) 对于所有的 $\theta\in\Theta$,集合 $\{x\mid f(x\mid\theta)>0\}$ 都是相同的.

(A2) 对数似然函数 $\ln f(x\mid\theta)$ 在区间 $[\theta_0-\delta,\theta_0+\delta]$ 关于 θ 三阶可导,$E_{\theta_0}l^{(1)}(\theta_0,X)$ 和 $E_{\theta_0}l^{(2)}(\theta_0,X)$ 都是有界的,且

$$\sup_{\theta\in(\theta_0-\delta,\theta_0+\delta)}\mid l^{(3)}(\theta,x)\mid\leqslant M(x),\ E_{\theta_0}M(x)<\infty,$$

其中 $h^{(i)}$ 代表函数 h 的第 i 阶导数.

(A3) 积分关于 θ 满足可交换性,$E_{\theta_0}l^{(1)}(\theta_0,X)=0$,$E_{\theta_0}l^{(2)}(\theta_0,X)=-E_{\theta_0}(l^{(1)}(\theta_0,X))^2$.并且 $I(\theta_0)=E_{\theta_0}(l^{(1)}(\theta_0,X))^2>0$ 成立.

(A4) 对于任意 $\delta>0$ 和足够大的 n,存在 $\epsilon>0$,使得 $\sup_{|\theta-\theta_0|>\delta}((l_n(\theta)-l_n(\theta_0))/n)<-\epsilon$ 依概率 1 成立.

(A5) 先验分布 $\pi_0(\theta)$ 是连续的,而且在 θ_0 处的密度大于 0.

证明 关于定理 1 的证明可以参考文献[32].

当参数 $\boldsymbol{\theta}$ 是一个低维度向量时,只需将上述正则性条件转化为相应的向量版本即可.由定理 1 可知:当样本量 n 增大时,后验分布 $\pi_n(\boldsymbol{\theta})$ 会越来越集中于真实参数 $\boldsymbol{\theta}_0$ 的附近,并且越来越接近于一个正态分布.

在一个实际问题中成功运用 Bayes 框架来完成数据分析,需要具备如下三个条件:① 要设定适当的先验分布;② 数据似然要具有明确的解析形式;③ 要能够较为便利地利用所得到的后验分布进行不确定性分析和推断.在先验分布的设定上,可以采用“主观先验”(subjective prior)来描述对未知参数的先验知识;也可以用“无信息先验”(noninformative prior),有时也被称作“客观先验”(objective prior),来表达对未知参数没有或者有很弱的先验知识.在统计学文献中,对相关的哲学逻辑、基本原则和实际方法有过大量的思辨和讨论[33-35].在利用后验分布进行推断方面,如果所得到的后验分布是常见的标准分布,则直接进行分析即可;如果得到的后验分布不是常见的标准分布,则可以通过分布近似或者 Monte-Carlo 抽样[36-39]的方式来进行处理.但对于数据似然具有明确的解析形式这一点,却是经典 Bayes 分析框架下的基本要求.这一方面给 Bayes 分析赋予了严格的概率含义,但同时也限制了 Bayes 分析在许多问题中的应用.在大数据和人工智能时代的许多实际问题中,由于数据和模型的复杂性,数据似然难以被精确刻画和计算的情况时有发生.研究如何在这种情形下利用生成模型进行有效的数据分析和不

确定性推断,具有重要的理论和现实意义.

1.2 经典 ABC 方法的基本框架和理论性质

在统计学文献中,ABC 方法是解决上述困境的一个重要方法.作为一种可以避免精确计算数据似然的抽样技术,ABC 方法的思想最早由 Rubin[22] 提出.后经过众多学者[23,40-45]的继承和发展,ABC 方法逐渐成为统计学习和 Bayes 计算的一个重要工具和活跃研究领域.ABC 方法的一个核心设定是:尽管真实数据生成过程 $f(\boldsymbol{y}_n \mid \boldsymbol{\theta})$ 的具体形式未知或者难以精确计算,但是对于任意给定的参数值 $\boldsymbol{\theta}$ 均可构造一个"数据生成器"(data generator) $G(\boldsymbol{z} \mid \boldsymbol{\theta})$,通过数值计算或者模拟的方式方便地从 $f(\boldsymbol{y}_n \mid \boldsymbol{\theta})$ 中产生新的样本.算法 1 给出了经典 ABC 方法的基本流程.

算法 1 ABC 算法

算法输入:

(I1) 观测数据 $\boldsymbol{y}_n$.

(I2) 模型参数 $\boldsymbol{\theta}$ 的先验分布 $\pi(\boldsymbol{\theta})$.

(I3) 数据生成器 $G(\boldsymbol{z} \mid \boldsymbol{\theta})$.

(I4) 观测数据统计量 $\eta(\boldsymbol{y}_n)$.

(I5) 观测数据统计量的距离度量 $D(\eta(\boldsymbol{z}),\eta(\boldsymbol{z}'))$.

(I6) 距离门限值 $\delta > 0$.

算法流程:

(S1) 从先验分布 $\pi(\boldsymbol{\theta})$ 中抽取一个候选参数 $\boldsymbol{\theta}^*$.

(S2) 从数据生成器 $G(\boldsymbol{z} \mid \boldsymbol{\theta}^*)$ 中产生一个合成数据 $\boldsymbol{z}$,并计算其统计量 $\eta(\boldsymbol{z})$.

(S3) 计算合成数据统计量 $\eta(\boldsymbol{z})$ 与观测数据统计量 $\eta(\boldsymbol{y}_n)$ 之间的距离 $D(\eta(\boldsymbol{z}),\eta(\boldsymbol{y}_n))$,如果 $D(\eta(\boldsymbol{z}),\eta(\boldsymbol{y}_n)) \leqslant \delta$ 则将 $\boldsymbol{\theta}^*$ 接受并保留,否则丢弃.

(S4) 重复步骤(S1)~(S3),直到算法的终止条件被满足.

算法输出:

(O1) 在算法运行过程中被保留下来的一组参数样本 $\boldsymbol{\theta}_1^*,\boldsymbol{\theta}_2^*,\cdots,\boldsymbol{\theta}_m^*$.

算法 1 的四个关键步骤为:先验分布抽样、构建数据生成器、计算统计量和计算统计量之间的距离.第一,从先验分布抽样.先验分布一般是标准分布,如正态分布、均匀分布等.因此,大多数软件包的内置函数都能轻易地从先验分布中产生样本.第二,构建数据生成器.数据生成器在不同的问题中是不同的.假设数据生成器是由一系列的常微分方程构成,那么在给定初始条件和参数值时,合成数据就可以通过数值模拟的方式产生.第三,计算统计量.如果数据存在低维度的充分统计量,那么我们就计算充分统计量.如果数据没有充分统计量,那么要计算一些概括统计量.概括统计量的选择需要研究者根据具体问题分析.第四,计算统计量之间的距离.理论上来说,我们可以计算观测数据统计量和合成数据统计量之间的任意被有效定义的距离,实际通常使用欧式距离.

算法 1 的操作流程将产生如下关于 $(\boldsymbol{\theta},\boldsymbol{z})$ 的联合分布:

$$\pi_\delta(\boldsymbol{\theta},\boldsymbol{z} \mid \boldsymbol{y}_n) = \frac{\pi_0(\boldsymbol{\theta}) f(\boldsymbol{z} \mid \boldsymbol{\theta}) I(\boldsymbol{z} \in \mathcal{A}_\delta)}{\int \pi_0(\boldsymbol{\theta}) f(\boldsymbol{z} \mid \boldsymbol{\theta}) I(\boldsymbol{z} \in \mathcal{A}_\delta) \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{\theta}}, \tag{3}$$

其中区域 $\mathcal{A}_\delta$ 定义为

$$\mathcal{A}_\delta = \{\boldsymbol{z}: D(\eta(\boldsymbol{z}),\eta(\boldsymbol{y}_n)) \leqslant \delta\}.$$

将 $\pi_\delta(\boldsymbol{\theta},\boldsymbol{z})$ 对 $\boldsymbol{z}$ 做积分,可得到关于 $\boldsymbol{\theta}$ 的一个边缘分布如下:

$$\pi_\delta(\boldsymbol{\theta}\mid \boldsymbol{y}_n)=\int\pi_\delta(\boldsymbol{\theta},z\mid \boldsymbol{y}_n)\mathrm{d}z=\frac{\pi_0(\boldsymbol{\theta})\int f(z\mid \boldsymbol{\theta})I(z\in\mathcal{A}_\delta)\mathrm{d}z}{\int\pi_0(\boldsymbol{\theta})f(z\mid \boldsymbol{\theta})I(z\in\mathcal{A}_\delta)\mathrm{d}z\mathrm{d}\boldsymbol{\theta}}. \tag{4}$$

称 $\pi_\delta(\boldsymbol{\theta}\mid \boldsymbol{y}_n)$ 为“ABC 后验分布”(ABC-posterior),并用它作为对真实后验分布 $\pi(\boldsymbol{\theta}\mid \boldsymbol{y}_n)$ 的近似.如下定理为 ABC 方法给出了基本的理论保证.

定理 2 如果 $\eta(\boldsymbol{y}_n)$ 是 $\boldsymbol{y}_n$ 的充分统计量且函数 $\eta(\cdot)$ 在 $\boldsymbol{y}_n$ 处连续,则在一定的正则性条件下:

$$\lim_{\delta\to 0}\int\mid\pi_\delta(\boldsymbol{\theta}\mid \boldsymbol{y}_n)-\pi(\boldsymbol{\theta}\mid \boldsymbol{y}_n)\mid\mathrm{d}\boldsymbol{\theta}=0.$$

证明 由于 $\eta(z)$ 是充分统计量,数据似然 $f(z\mid\boldsymbol{\theta})$ 存在如下分解:

$$f(z\mid\boldsymbol{\theta})=g(\eta(z),\boldsymbol{\theta})\cdot h(z),\qquad \forall(z,\boldsymbol{\theta}).$$

因而

$$\pi(\boldsymbol{\theta}\mid \boldsymbol{y}_n)\propto\pi_0(\boldsymbol{\theta})g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta}), \tag{5}$$

$$\pi_\delta(\boldsymbol{\theta}\mid \boldsymbol{y}_n)\propto\pi_0(\boldsymbol{\theta})\int g(\eta(z),\boldsymbol{\theta})h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z. \tag{6}$$

定义

$$s(\boldsymbol{\theta})=g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta}),$$

$$s_\delta(\boldsymbol{\theta})=\int g(\eta(z),\boldsymbol{\theta})h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z,$$

$$s_\delta^*(\boldsymbol{\theta})=\int g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta})h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z,$$

并令 $C=\int\pi_0(\boldsymbol{\theta})s(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, C_\delta=\int\pi_0(\boldsymbol{\theta})s_\delta(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, C_\delta^*=\int\pi_0(\boldsymbol{\theta})s_\delta^*(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$, 有

$$\pi(\boldsymbol{\theta}\mid \boldsymbol{y}_n)=\pi_0(\boldsymbol{\theta})s(\boldsymbol{\theta})/C,$$

$$\pi_\delta(\boldsymbol{\theta}\mid \boldsymbol{y}_n)=\pi_0(\boldsymbol{\theta})s_\delta(\boldsymbol{\theta})/C_\delta,$$

且容易验证

$$s_\delta^*(\boldsymbol{\theta})/C_\delta^*=s(\boldsymbol{\theta})/C.$$

进而有

$$\int\mid\pi_\delta(\boldsymbol{\theta}\mid \boldsymbol{y}_n)-\pi(\boldsymbol{\theta}\mid \boldsymbol{y}_n)\mid\mathrm{d}\boldsymbol{\theta}=$$

$$\int\left|\frac{s_\delta(\boldsymbol{\theta})}{C_\delta}-\frac{s(\boldsymbol{\theta})}{C}\right|\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}\leqslant$$

$$\int\left|\frac{s_\delta(\boldsymbol{\theta})}{C_\delta}-\frac{s_\delta^*(\boldsymbol{\theta})}{C_\delta}\right|\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}+\int\left|\frac{s_\delta^*(\boldsymbol{\theta})}{C_\delta}-\frac{s(\boldsymbol{\theta})}{C}\right|\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}=$$

$$\int\left|\frac{s_\delta(\boldsymbol{\theta})-s_\delta^*(\boldsymbol{\theta})}{C_\delta}\right|\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}+\frac{\mid C_\delta-C_\delta^*\mid}{C_\delta}\leqslant$$

$$2\int\left|\frac{s_\delta(\boldsymbol{\theta})-s_\delta^*(\boldsymbol{\theta})}{C_\delta}\right|\pi_0(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}.$$

易知

$$\left|\frac{s_\delta(\boldsymbol{\theta})-s_\delta^*(\boldsymbol{\theta})}{C_\delta}\right|\leqslant\frac{1}{C_\delta}\int\mid g(\eta(z),\boldsymbol{\theta})-g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta})\mid h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z\leqslant$$

$$\sup_{z\in\mathcal{A}_\delta}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|g(\eta(z),\boldsymbol{\theta})-g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta})\right|\cdot\frac{\int h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z}{C_\delta},$$

其中

$$\lim_{\delta\to 0}\frac{\int h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z}{C_\delta}=$$

$$\lim_{\delta\to 0}\frac{\int h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z}{\int\pi_0(\boldsymbol{\theta})g(\eta(z),\boldsymbol{\theta})h(z)I(z\in\mathcal{A}_\delta)\mathrm{d}z\mathrm{d}\boldsymbol{\theta}}=$$

$$\frac{1}{\int\pi_0(\boldsymbol{\theta})g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}}\triangleq M.$$

若函数 $g(\eta,\boldsymbol{\theta})$ 一致连续,则对于 $\epsilon>0$,存在 $\delta>0$, 使得

$$\sup_{z\in\mathcal{A}_\delta}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|g(\eta(z),\boldsymbol{\theta})-g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta})\right|\leqslant\epsilon,$$

即

$$\lim_{\delta\to 0}\sup_{z\in\mathcal{A}_\delta}\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|g(\eta(z),\boldsymbol{\theta})-g(\eta(\boldsymbol{y}_n),\boldsymbol{\theta})\right|=0.$$

进而,定理 2 得证. □

上述定理表明:只要 ABC 算法中选取的统计量 η 是充分统计量,在一定的正则性条件下,随着距离门限值 δ 趋近于 0, ABC 后验分布 $\pi_\delta(\boldsymbol{\theta}\mid\boldsymbol{y}_n)$ 会收敛到真实的后验分布 $\pi(\boldsymbol{\theta}\mid\boldsymbol{y}_n)$,即 ABC 后验对真实后验的近似误差会趋近于 0.从上述定理的证明过程中还可以看到:若 η 不是充分统计量,则 ABC 后验对真实后验的近似会产生不可消除的理论误差,误差的大小取决于将数据 z 转化为统计量 $\eta(z)$ 的过程中所产生的信息损失.同时,易知 ABC 方法中接受一个候选参数 $\boldsymbol{\theta}^*$ 的概率(即“接受率”,acceptance rate)为

$$C_\delta=\int\pi_0(\boldsymbol{\theta})f(z\mid\boldsymbol{\theta})I(D(\eta(z),\eta(\boldsymbol{y}_n))\leqslant\delta)\mathrm{d}z\mathrm{d}\boldsymbol{\theta}.$$

显然,随着距离门限值 δ 趋近于 0,有 $C_\delta\to 0$, 即 ABC 方法的计算负担会逐渐增大直到实际的计算效率降低为 0.在实际运用中,为了保持适当的计算效率,必须要选取一个非零的门限值 δ, 从而会不可避免地在 ABC 后验中引入计算误差,误差的大小取决于接受区域 $\{z: D(\eta(z),\eta(\boldsymbol{y}_n))<\delta\}$ 的几何特性.由于接受区域 $\{z: D(\eta(z),\eta(\boldsymbol{y}_n))<\delta\}$ 由统计量 η、距离 D 和门限值 δ 共同决定,ABC 后验分布的总误差(包括理论误差和计算误差)本质上由上述三个要素共同决定.更多关于 ABC 方法理论性质的讨论可以参考文献[45-46].

1.3 经典 ABC 方法的实际运用

由于使用了数据统计量之间相似度的比较来替代似然函数的精确计算,ABC 方法放松了 Bayes 推断中似然函数需要有解析形式的限制条件,从而极大地扩展了 Bayes 推断框架的适用范围,使之可以在似然函数没有显式的解析式或者难以精确计算的情况下仍能有效运行.

早期运用 ABC 方法一个经典的例子是 1999 年 Pritchard 等关于群体遗传学(population genetics)的一项研究[23].在该研究中,观测到的数据集 $\boldsymbol{y}_n$ 是一些个体的基因序列,研究目标是使用这些观测数据推断群体遗传的一些关键参数,比如基因突变率等.解决这类问题的一个常用模型是 Kingman's coalescent 模型[47],其数据似然的形式如下所示:

$$f(\boldsymbol{y}_n \mid \boldsymbol{\theta}) = \int_{\boldsymbol{H}} f(\boldsymbol{y}_n \mid \boldsymbol{H}) f(\boldsymbol{H} \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{H}, \tag{7}$$

其中 $\boldsymbol{H}$ 代表在总体进化过程中未观测到的系谱历史事件集合.典型的 $\boldsymbol{H}$ 是一个元素个数可变、事件类型未知的高维集合.因此,上述似然函数一般而言没有显式表达式.类似的例子在传染病学[48]和生态学[49]研究中广泛存在.

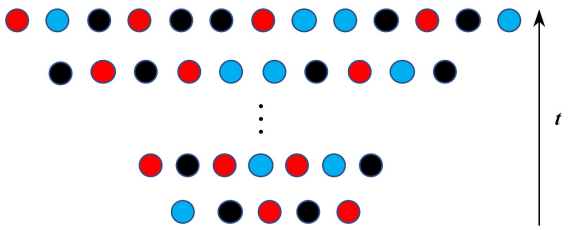

图 1 总体数量变化的 coalescence 模型

Fig. 1 The coalescence model with population size growth

图 1 展示了 coalescence 模型,图中每行代表一个时代,不同的颜色代表不同的显型,代际之间会发生各种事件,包含死亡、生产、迁徙等,最上面一行代表当前时代的总体显型分布.虽然式(7)没有显式表达式,但是给定参数时生成这类的数据是相对容易做到的.因此,ABC 方法可以处理这类问题.

另外一类有代表性的例子是空间统计中常用的 Potts 模型[50].图 2 展示了一个定义在 4×4 格点上的 Potts 模型.广义 Potts 模型是对分布在空间格点上的数据进行空间相关性建模的一种重要方法,在统计物理、图像处理等领域有广泛应用.给定 d 维空间中特定区域内的一组格点 $\mathcal{L}$,假设每个格点位置 i 对应于一个取值范围为 $\Omega = \{1,2,\cdots,q\}$ 的随机变量 y_i,并记 $\boldsymbol{y} = \{y_i\}_{i\in\mathcal{L}}$.对于格点结构中的任意两个位置 i, j,如果位置 i 与位置 j 相邻,则记作 $i \sim j$.Potts 模型假设 $\boldsymbol{y}$ 具有如下形式的数据似然:

$$f(\boldsymbol{y} \mid \theta) = \frac{\exp\left\{-\theta \sum_{i\sim j} I(y_i = y_j)\right\}}{\mathcal{C}(\theta)}, \tag{8}$$

其中正则化常数 $\mathcal{C}(\theta)$ 的表达式为

$$\mathcal{C}(\theta) = \sum_{\boldsymbol{z}\in\Omega^{\#\mathcal{L}}} \exp\left\{-\theta \sum_{i\sim j} I(z_i = z_j)\right\}.$$

显然,对 Potts 模型数据似然的精确计算涉及到对正则常数 $\mathcal{C}(\theta)$ 的计算,这需要遍历 $\boldsymbol{z}$ 所有可能的取值状态,其计算复杂度为 $\mathcal{O}(q^{\#\mathcal{L}})$.因而,即使对于规模较小的 Potts 模型,其似然函数的精确计算都是非常困难的.实际上,Potts 模型是 Gibbs 随机域(Gibbs random field,GRF)的一个特例,而几乎所有关于 GRF 的数据分析问题都会遇到类似的计算挑战,使运用 Bayes 分析面临较大的计算困难.

给定参数 θ,有多种算法可以生成相应的 Potts 样本.同时,Potts 模型的充分统计量是 $\sum_{i\sim j} I(z_i = z_j)$.这样,ABC 方法可以避免计算正则化常数而对参数进行 Bayes 推断.

1.4 算法配置

基于前述分析, ABC 后验分布的误差由统计量 η、距离 D 和门限值 δ 这三个要素共同决定.因而,在一个实际问题中运用 ABC 方法时,我们应在 ABC 方法的框架下对上述三个要素进行合理配置,以期在近似的精确性和计算的复杂性之间达到平衡,以合理的计算代价获得对真

实后验分布适当的近似.从统计学一般原理上讲,统计量 η 应该尽可能地选取极小充分统计量(minimal sufficient statistics)为好.这是因为极小充分统计量能够在保持充分性的同时最大程度地实现对数据的压缩和降维,进而方便后续距离度量的设计和计算,从而提高算法的运行效率.然而,由于数据生成模型 $G(\boldsymbol{y}\mid\boldsymbol{\theta})$ 的具体形式一般未知,确定极小充分统计量常常并不现实.这时,就需要算法的设计者根据实际问题的具体背景和专业知识来选取近似满足极小充分性的统计量.在经典 ABC 方法的框架下,统计量 η 的选取常常对整个算法的效率具有决定性的影响.在距离 D 的选择上,除了最常用的欧氏距离之外,还可以选择 Manhattan 距离、马氏距离等多种不同的距离.在不同的实际问题中,不同距离度量下的计算效率会有差异.如何选择适当的距离以达到较高的计算效率是 ABC 方法研究中的一个热点问题.接受域半径 δ 的选择是一个平衡计算成本和近似精度的过程,可以根据具体问题的应用场景和计算资源的情况进行配置.最后,我们还需要为算法设定明确的终止条件.常见的终止条件有两种:根据重复步骤(S1)~(S4)的总次数终止算法,或者根据获得的后验样本数量终止算法.但由于算法步骤(S3)中接受候选参数 $\boldsymbol{\theta}^*$ 的比率 r 是一个恒定的常数,这两种终止条件本质上具有等价性.

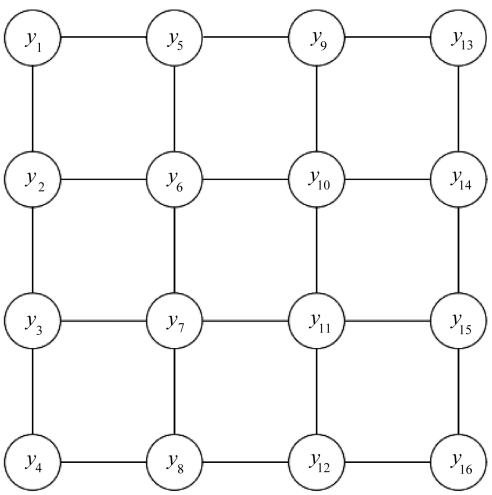

图 2 定义在 4×4 格点上的 Potts 模型

Fig. 2 A Potts model defined on the 4×4 lattice

2 ABC 方法前沿进展

近年来,众多研究者从不同的角度对经典 ABC 方法进行了改进,形成了一系列方法.本节将对这些前沿研究进展进行综述和讨论.

2.1 以数据驱动的方式筛选统计量

如前所述,选取适当的统计量是运用 ABC 方法的一个重要步骤.在经典的 ABC 框架下,统计量的选取主要依靠算法的设计者依据对实际问题的理解和经验进行人为设计.但是,在很多实际问题中,这种设计并不总是容易的.很多时候,一种统计量的选择并不显然比另一种好.而且,有研究表明统计量的不同选择会对 ABC 方法的实际表现产生较大的影响[51].为了克服这种困境,运用数据驱动的方式自动筛选统计量的策略近年来越来越受到重视.

Joyce 和 Marjoram 提出一种近似充分统计量的方法[52].首先,统计量的集合 S 是一组试验性的统计量集合.集合 S 尽可能多地包含统计量.他们认为,从 S 中挑选一部分统计量 $\boldsymbol{s}$ 就可以达到 S 的效果.从数学角度而言,如果 $\pi(\boldsymbol{\theta}\mid s_1,s_2,\cdots,s_k)/\pi(\boldsymbol{\theta}\mid s_1,s_2,\cdots,s_{k-1})-1\leqslant\delta$ 成立,

s_k 会被认为独立于参数 $\boldsymbol{\theta}$,因此 s_k 很可能不会被纳入集合 $\boldsymbol{s}$.其中 δ 是研究者自定义的一个数值很小的阈值.选择 $\boldsymbol{s}$ 的具体方法请参考文献[52].这种方法的缺陷也很明显,对于大部分需要使用 ABC 的模型而言,统计量的数量并不存在一个上限.他们的研究表明,即使模型相同,当数据不同时 $\boldsymbol{s}$ 也会改变.

Fearnhead 和 Prangle 提出半自动化 ABC(semi-automatic ABC)来选择统计量[42].他们首先定义如下的损失函数:

$$L = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{A}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}), \tag{9}$$

其中 $\boldsymbol{\theta}$ 为真实值,$\hat{\boldsymbol{\theta}}$ 是估计值.在选择合适的矩阵 $\boldsymbol{A}$ 时,当统计量 $\boldsymbol{s} = E(\boldsymbol{\theta} \mid \boldsymbol{y})$ 时,L 会被最小化.换句话说,参数的后验分布均值是基于式(9)定义的损失函数下的最优统计量.此时,我们陷入两难的困境,因为不知道真实后验分布,后验分布的均值自然也是未知的.尽管如此,真实后验分布均值的估计量依然是可供我们选择的统计量.半自动化 ABC 的过程可以由算法 2 描述如下.

算法 2 半自动化 ABC 算法

1) 使用算法 1 做试验性估计.确定参数的大致取值区间.

2) 在上述取值区间中抽取候选参数,并根据似然函数生成合成数据.重复 M 次该步骤,因此得到 $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_M$ 和 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_M$.

3) 使用 $\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\cdots,\boldsymbol{\theta}_M$ 和 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_M$ 分别估计数据和后验分布均值之间的关系.其中,第 $i \in \{1,2,\cdots,p\}$ 个参数 θ_i 是因变量,$\boldsymbol{z}$ 是自变量.即,$\theta_i = \hat{F}_i(\boldsymbol{z})$.

4) 给定统计量 $\hat{F}_1(\cdot),\hat{F}_2(\cdot),\cdots,\hat{F}_p(\cdot)$,执行 ABC 算法 1,得到后验样本.

半自动化 ABC 算法的关键在于步骤 3).Fearnhead 和 Prangle 实验了多种方法来拟合 $\boldsymbol{\theta}$ 与数据 $\boldsymbol{y}_n$ 之间的关系,包括 lasso[53] 和 canonical correlation analysis[54].他们发现,上述方法并没有比简单线性回归表现得更好.因此,步骤 3)中使用下面的模型拟合第 i 个参数与数据之间的关系:

$$\theta_i = \beta_0 + \boldsymbol{\beta}\boldsymbol{f}(\boldsymbol{z}) + \epsilon, \tag{10}$$

其中 $\boldsymbol{f}(\boldsymbol{z})$ 可以是 $\boldsymbol{z}$ 的任意组合.Fearnhead 和 Prangle 采用 $\boldsymbol{f}(\boldsymbol{z}) = (\boldsymbol{z},\boldsymbol{z}^2,\boldsymbol{z}^3,\boldsymbol{z}^4)$.

在 Fearnhead 和 Prangle 提出半自动化 ABC 后,有学者扩展了步骤 3)中的函数关系.Jiang 等使用深度神经网络来拟合统计量与合成数据之间的关系[55].理论上来说,深度神经网络可以更好地拟合它们之间的函数关系.但是深层的神经网络需要大量的数据来学习参数.所以,Jiang 等使用三隐藏层的神经网络来执行模型.使用神经网络的研究者还包括 Creel[56].Raynal 等提出使用随机森林的方法选择统计量[57].总体而言,目前的方法大部分集中于寻找数据与统计量之间的函数关系.这类方法比较依赖于额外的训练数据,即函数关系需要通过额外合成数据来训练得到.更多关于选择统计量的研究综述可以参考文献[58].

2.1.1 例 1:"g-and-k" 分位数分布(quantile 'g-and-k' distribution)

一元 "g-and-k" 分位数分布是 ABC 领域中常见的用于测试算法的模型[42,44].该分布由如下的形式定义:

$$Q(r \mid A,B,g,k,c) = A + B\left(1 + c\,\frac{1 - \exp(-gz(r))}{1 + \exp(-gz(r))}\right)(1 + z(r)^2)^k z(r), \tag{11}$$

其中,$z(r)$ 表示标准正态分布的第 r 个分位数,参数 c 衡量分布的总体非对称性,一般设定为 0.8,需要推断的参数包括 A,B,g,k.本小节对比两种算法的表现:算法 1 和算法 2.

给定参数的真实值 $(A^*,B^*,g^*,k^*) = (3,1,2,0.5)$.数据根据真实值生成,其中样本量

为 $n = 250$.因为“g-and-k”分位数分布并没有已知的充分统计量.我们使用数据的一系列分位数作为数据集的统计量.在算法中 η 被选定为数据的一系列分位数.其中,分位数点为[0.05,0.1,0.15,…,0.9,0.95].算法 1 的距离度量选择欧氏距离.对于两种算法,每种算法生成 10^6 次合成数据.算法 1 选择其中距离最小的 $N = 2\ 048$ 个候选参数作为其后验样本.同样地,算法 2 也产生 N 个后验样本.半自动化 ABC 的代码发布在 R 软件中的 abctools 包中.本文使用 abctools 包来实现参数推断.

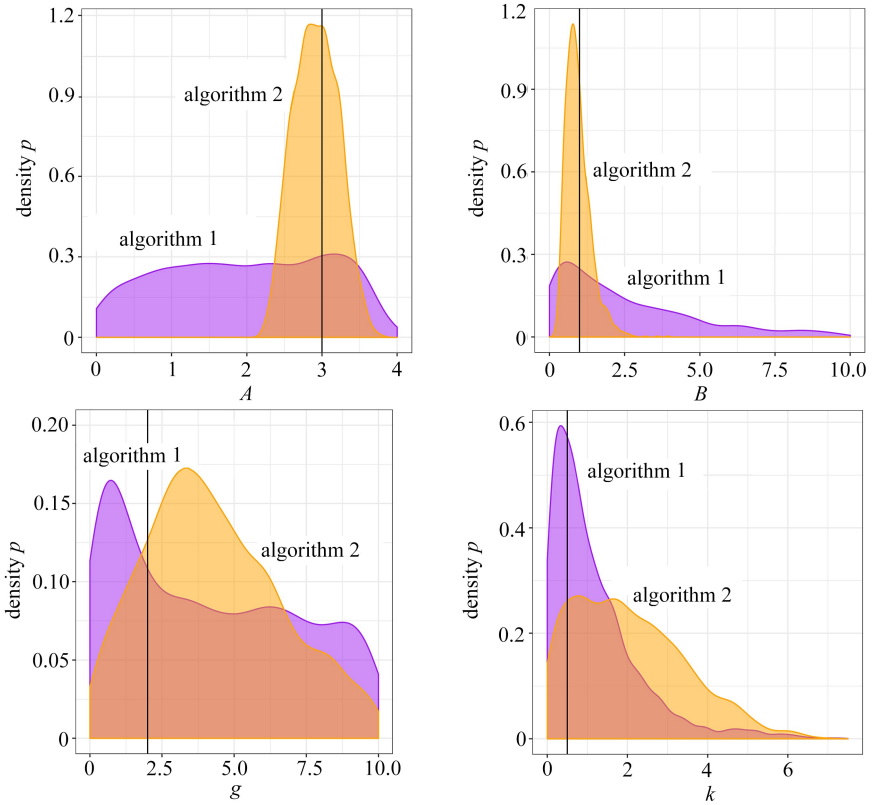

图 3 对比算法 1 和算法 2 得到的参数 (A,B,g,k) 的后验样本边际分布

Fig. 3 Comparison of marginal distributions of (A,B,g,k) between algorithm 1 and 2

图 3 展示了算法 1 和 2 得到的后验样本结果,其中深色代表算法 1 的结果,浅色代表算法 2 的结果,每个子图中的黑色竖线代表每个参数的真实值.从图中可以看出,算法 1 在参数 k 的表现上是优于半自动化 ABC 的.在参数 g 上两者的表现相当,半自动化 ABC 略微优于算法 1.在参数 A,B 上半自动化 ABC 的表现明显好过算法 1.所以,总体上来看半自动化 ABC 表现胜过算法 1,而且半自动化 ABC 并不需要使用者主动设计统计量.

2.2 采用更有效的距离度量

优化选取统计量之间的距离度量是改进 ABC 方法的另外一个重要途径.多位研究者意识到统计量的尺度差异会影响 ABC 的表现[45,49].Harrison 等是为数不多的研究如何消除统计量之间的尺度差异影响的研究者[59],文献中以最常见的欧氏距离为例做研究.通常,经过统计量降维后,观测数据与合成数据之间的距离可以表示为

$$D(\eta(\boldsymbol{y}),\eta(\boldsymbol{z})) = \sqrt{\sum_{k=1}^{K}(\eta_k(\boldsymbol{y}) - \eta_k(\boldsymbol{z}))^2},$$

其中 K 是统计量的个数.为了消除统计量尺度大小的影响,Harrison 等引入加权欧氏距离.其定义如下:

$$D_{\boldsymbol{w}}(\eta(\boldsymbol{y}),\eta(\boldsymbol{z}))=\sqrt{\sum_{k=1}^{K}\boldsymbol{w}_k(\eta_k(\boldsymbol{y})-\eta_k(\boldsymbol{z}))^2}, \tag{12}$$

其中 $\boldsymbol{w}$ 代表统计量的权重向量.权重向量的选择会影响 ABC 算法得到的后验分布.文献[59]的目标就是选择最优的权重向量 $\boldsymbol{w}^*$ 使得参数的先验分布与 ABC 后验分布之间的距离最大.在文献[59]中,Hellinger 距离被用来衡量两个分布之间的距离.当然,其他距离度量,如欧氏距离、Kullback-Leibler 散度都可以作为备选.在给定两个分布的样本时,最近邻距离估计量(nearest neighbour distance estimator)被用来估计分布之间的距离.两组样本集之间的 k- 最近邻距离估计的定义如下:

$$\hat{D}_{\alpha}(X_{1:n}\parallel Y_{1:n})=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(n-1)\rho_k(i)}{n\nu_k(i)}\right)^{1-\alpha}B_{k,\alpha}, \tag{13}$$

其中 $\rho_k(i)$ 指 X_i 到它在 $X_{1:n}$ 中第 k 近样本的欧氏距离,$\nu_k(i)$ 指 X_i 到它在 $Y_{1:n}$ 中第 k 近样本的欧氏距离,$B_{k,\alpha}=\Gamma(k)^2/(\Gamma(k-\alpha+1)\Gamma(k+\alpha-1))$.因此,最优的权重向量可以由最优化来解决:

$$\boldsymbol{w}^*=\arg\max_{\boldsymbol{w}}(1-\hat{D}_{\alpha}(\boldsymbol{\theta}'_{1:n}\parallel\boldsymbol{\theta}_{1:n})), \tag{14}$$

其中 $\alpha=0.5$,$\boldsymbol{\theta}'_{1:n}$是从先验分布中抽取的样本,$\boldsymbol{\theta}_{1:n}$ 是 ABC 算法得到的后验分布样本.需要注意的是 $\boldsymbol{\theta}_{1:n}$ 与 $\boldsymbol{w}$ 相关.

由于观测数据的可交换性,直接使用数据集作为统计量在很多时候并不可行[58].但是最近有学者指出数据的可交换性并不会成为一个问题,并提出相应的方案实施 ABC 算法.比较典型的研究包括 Bernton 等[60]提出使用 Wasserstein 距离作为 ABC 算法中的距离度量.这样做有两个优势:首先,Wasserstein 距离可以直接作用在数据集本身,它不需要统计量降维,因此就避免选择统计量的工作.第二,由 Wasserstein 距离诱导的 ABC 算法可以收敛到真实的后验分布,其原因在于数据集本身是一个充分统计量.算法 3 是采用 Wasserstein 距离的 ABC 算法流程,以下简称 WABC.

算法 3 WABC 算法

1) 从先验分布 $\boldsymbol{\theta}$ 中抽取一个候选参数 $\boldsymbol{\theta}^*$.

2) 给定 $\boldsymbol{\theta}^*$,根据真实数据生成过程产生一个合成数据 $\boldsymbol{z}$.

3) 计算 $W(\boldsymbol{z},\boldsymbol{y})$,其中 $W(\cdot,\cdot)$ 代表 Wasserstein 距离.

4) 如果 $W(\boldsymbol{z},\boldsymbol{y})\leqslant\delta$,保留 $\boldsymbol{\theta}^*$ 和 $\boldsymbol{z}$.

5) 重复步骤 1)~4),直到算法的终止条件被满足.

2.2.1 例 2:回访“g-and-k”分位数分布

我们回访在第 2.1.1 小节中的例子,使用算法 3 对观测数据做参数推断.算法 3 的代码可以在 github.com/pierrejacob/winference 找到.为了寻找一个合理的阈值 ϵ,WABC 与序列 Monte-Carlo 方法结合.关于序列 Monte-Carlo 的介绍,可以参考本文第 2.3.2 小节.

图 4 展示了算法 2 和算法 3 得到的后验样本的边际分布情况,其中深色代表算法 3 的结果,浅色代表算法 2 的结果,每个子图中的黑色竖线代表每个参数的真实值.很显然,使用 Wasserstein 距离的 ABC 算法明显优于半自动化 ABC.在每个参数的边际分布上,算法 3 都是更优的.不仅如此,WABC 还极大地简化了使用 ABC 算法的过程.使用者不需要设计统计量来降低

数据的维度.但是 WABC 遇到的另一个问题是 Wasserstein 距离的计算复杂度高.如果观测样本的样本量为 n 且样本是多元数据,那么精确 Wasserstein 距离的计算复杂度为 $\mathcal{O}(n^3)$.在文献[60]中,研究者提供了关于近似求解 Wasserstein 距离的一些方法.它们都可以降低计算成本,但是需要牺牲一部分计算精度.

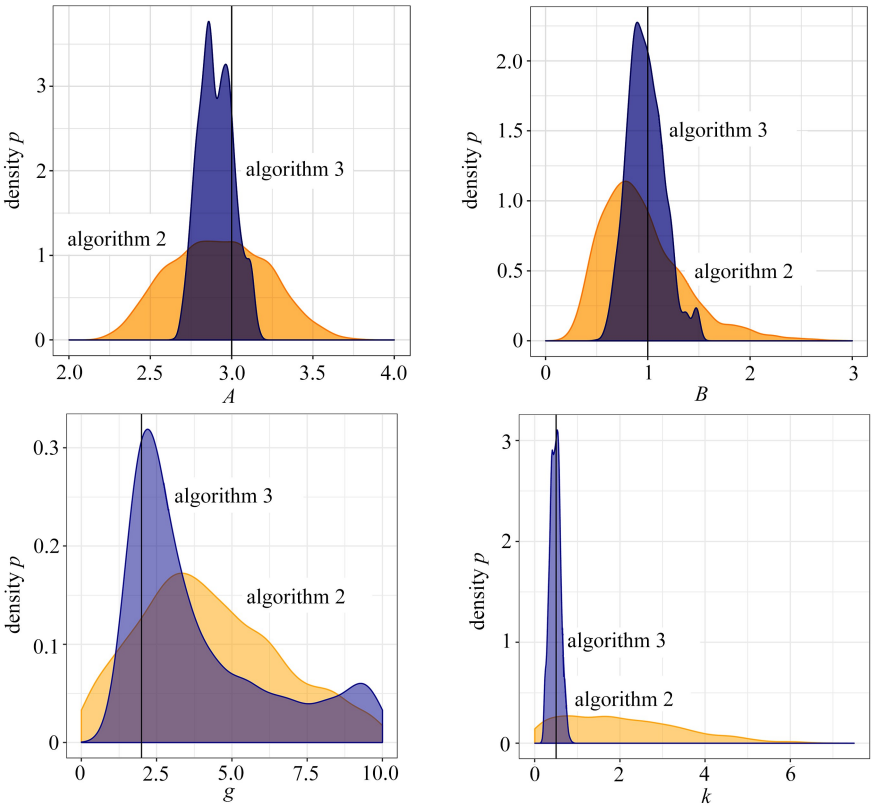

图 4 对比算法 2 和算法 3 得到的参数 (A,B,g,k) 的后验样本边际分布

Fig. 4 Comparison of marginal distributions of (A,B,g,k) between algorithm 2 and 3

2.3 计算框架的进一步扩展

实际上,经典 ABC 算法框架的计算效率是比较低的.为了提高 ABC 算法的计算效率,学者们对经典 ABC 算法的计算框架进行了一系列扩展和改进,将 ABC 方法与其他的抽样和分析技术进行有效整合以提高计算效率.

2.3.1 Soft ABC

算法 1 使用拒绝和接受两种极端的方式对待候选参数.这种方式被称为“拒绝抽样 ABC”.与之对应的是 soft ABC,其具体流程如算法 4 所示.

算法 4 Soft ABC 算法

1) 从先验分布 $\pi(\boldsymbol{\theta})$ 中抽取一个候选参数 $\boldsymbol{\theta}^*$.

2) 给定 $\boldsymbol{\theta}^*$,根据真实数据生成过程产生一个合成数据 $\boldsymbol{z}$.

3) 计算相似度 $r=\exp(-D(\boldsymbol{z},\boldsymbol{y}_n)/h)$,其中 $D(\cdot,\cdot)$ 代表某种距离度量(如,欧氏距离).

4) 依概率 r 保留 $\boldsymbol{\theta}^*$ 和 $\boldsymbol{z}$.

5) 重复步骤 1)~4),直到算法的终止条件被满足.

显然,算法 4 中 h 的作用与算法 1 中的 δ 类似,控制了 soft ABC 方法的近似精度和计算效

率.当 $h \to 0$ 时,soft ABC 方法的近似精度逐渐提高,但计算效率逐渐降低.在实际应用 soft ABC 的过程中,我们需要适当地选择 h.

在此基础上,Park 等在 2016 年提出可以用两个样本之间的最大均值差异(maximum mean discrepancy,MMD)替代其距离度量 D,对 soft ABC 方法进行改造.该算法被简称为 K2-ABC[61],其具体流程如算法 5 所示.

算法 5 K2-ABC 算法

1) 从先验分布 $\pi(\boldsymbol{\theta})$ 中抽取一个候选参数 $\boldsymbol{\theta}^*$.

2) 给定 $\boldsymbol{\theta}^*$,根据真实数据生成过程产生一个合成数据 $\boldsymbol{z}$.

3) 计算相似度 $\delta = \exp(-D_{\mathrm{MMD}}(\boldsymbol{z},\boldsymbol{y}_n)/h)$,其中 D_{MMD} 表示距离.

4) 依概率 δ 保留 $\boldsymbol{\theta}^*$ 和 $\boldsymbol{z}$.

5) 重复步骤 1)~4),直到算法的终止条件被满足.

2.3.2 ABC-SMC

经典 ABC 框架的一大局限在于:候选参数是从参数的先验分布抽取的,当先验分布与后验分布差异较大时,一般会造成很高的抽样拒绝率.这一弊端可以通过巧妙地构造更加接近于后验分布的抽样分布来部分克服.ABC-SMC 算法[41,62]对经典 ABC 框架中产生候选参数的抽样过程进行了改造,通过序列 Monte-Carlo 策略逐步优化更新候选参数的抽样分布.ABC-SMC 算法由 $T+1$ 个串行的 ABC 过程所构成,每一个 ABC 过程对应于一个不同的阈值 δ_t,其中 $\infty = \delta_0 > \delta_1 > \cdots > \delta_T$.其详细流程如算法 6 所示.

算法 6 ABC-SMC 算法

1) 令 $t=0$.从先验分布 $\pi_0(\boldsymbol{\theta})$ 中抽取 N 个样本 $\{\boldsymbol{\theta}_i^{(0)}\}_{i=1}^N$,并对每个样本赋予权重 $w_i^{(0)} = 1/N$.

2) 令 $t=t+1$.构造混合分布 $G^{(t)}(\boldsymbol{\theta}) = \sum_{i=1}^N w_i^{(t-1)} K(\boldsymbol{\theta} \mid \boldsymbol{\theta}_i^{(t-1)})$,其中 $K(\cdot \mid \cdot)$ 是给定的转移核函数.

3) 令 $i=1$.

4) 以混合分布 $G^{(t)}(\boldsymbol{\theta})$ 为抽样分布,以 δ_t 为门限值,运行 ABC 算法获得关于参数 $\boldsymbol{\theta}$ 的新样本:首先从 $G^{(t)}(\boldsymbol{\theta})$ 中抽样得到 $\boldsymbol{\theta}_i^{(t)}$,进而从数据生成器生成合成数据 $\boldsymbol{z} \sim f(\cdot \mid \boldsymbol{\theta}_i^{(t)})$,并重复该步骤直到 $D(\eta(\boldsymbol{y}_n),\eta(\boldsymbol{z})) \leqslant \delta_t$.

5) 令 $i=i+1$.返回步骤 4),直至 $i > N$.

6) 赋予样本权重:

$$w_i^{(t)} = \frac{\pi(\boldsymbol{\theta}_i^{(t)})}{\sum_{j=1}^N w_j^{(t-1)} K(\boldsymbol{\theta}_i^{(t)} \mid \boldsymbol{\theta}_j^{(t-1)})}.$$

7) 返回步骤 2),直至 $t > T$.

在算法 6 中转移核函数 $K(\cdot \mid \cdot)$ 通常可以选择 Gauss 核函数.由于在 t 时刻,候选参数不再是从先验分布 $\pi_0(\boldsymbol{\theta})$ 抽取,而是从以 $t-1$ 时刻获得的样本为局部中心所形成的 Gauss 混合分布 $G^{(t)}(\boldsymbol{\theta})$ 抽取,ABC-SMC 算法可以有效继承我们在 t 时刻之前对后验分布的近似认知,从而提高算法的抽样效率.在 ABC 中使用 SMC 方法的研究还包括文献[63-64].

2.3.3 ABC-MCMC

Marjoram 等提出了在 ABC 框架中使用 MCMC 的方法以提高计算效率[40].具体算法流程

如算法 7 所示.在算法 7 中,似然函数不需要被计算.同时,当 $\delta \to 0$ 时,算法得到的后验分布收敛到真实的后验分布.ABC-MCMC 的优势在于,候选参数不需要从先验分布获得.通常 $K(\cdot \mid \cdot)$ 选择 Gauss 核函数.如果 $K(\cdot \mid \cdot)$ 和参数的先验分布都选择均匀分布,那么 ABC-MCMC 就会退化为算法 1.如果目标分布是多峰分布,ABC-MCMC 面临的挑战是后验样本容易陷入局部单峰区间.

算法 7 ABC-MCMC 算法

1) 令 $t = 0$.使用算法 1 得到一个后验样本 $\boldsymbol{\theta}^{(0)}$.

2) 令 $t = t + 1$.

3) 抽样候选参数 $\boldsymbol{\theta}^* \sim K(\cdot \mid \boldsymbol{\theta}^{(t-1)})$.

4) 给定 $\boldsymbol{\theta}^*$,根据真实数据生成过程产生一个合成数据 $\boldsymbol{z}$.

5) 生成随机数 $u \sim U(0,1)$.

6) 计算 $r = \dfrac{\pi(\boldsymbol{\theta}^*)K(\boldsymbol{\theta}^{(t-1)} \mid \boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^{(t-1)})K(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(t-1)})}$.

7) 如果 $u \leqslant r$ 和 $D(\eta(\boldsymbol{y}_n),\eta(\boldsymbol{z})) \leqslant \delta$ 都成立,则令 $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$,$\boldsymbol{z}^{(t)} = \boldsymbol{z}$.否则,$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

8) 返回步骤 2),直到算法的终止条件被满足.

2.3.4 Hamiltonian ABC

Hamiltonian Monte-Carlo(HMC)方法是一种常见的处理高维推断的方法[65].HMC 利用目标函数的一些信息(如,梯度信息和 Hesse 矩阵等)来提高参数的抽样效率.但是 ABC 不需要任何密度函数的信息.直觉上来说,HMC 与 ABC 是两种思路相反的框架.但是 Meeds 等将两种方法结合[66],这种方法被简称为 HABC. HABC 可以同时拥有 HMC 和 ABC 的优势:处理高维参数推断和免于计算似然函数.

面对复杂模型,似然函数难以计算.HMC 缺少似然函数的信息就不能进行抽样.但是 ABC 得到的后验样本却可以帮助我们了解似然函数的大致信息.在 HABC 中一个关键的公式是

$$\pi_\delta(\boldsymbol{y}_n \mid \boldsymbol{\theta}) = \int \pi_\delta(\boldsymbol{y}_n \mid \boldsymbol{z})\pi(\boldsymbol{z} \mid \boldsymbol{\theta})\mathrm{d}\boldsymbol{z} \approx \frac{1}{M}\sum_{m=1}^{M}\pi_\delta(\boldsymbol{y}_n \mid \boldsymbol{z}^{(m)}), \tag{15}$$

其中 $\boldsymbol{z}^{(1)},\boldsymbol{z}^{(2)},\cdots,\boldsymbol{z}^{(M)}$ 是由 ABC 算法获得的 M 个后验样本,$\pi_\epsilon(\boldsymbol{y}_n \mid \boldsymbol{z})$ 表示合成数据与观测数据之间的差异,也可以当作似然函数.因此,对于给定的 $\boldsymbol{\theta}$, 观测数据的似然函数都可以根据式(15)计算. 相应地, 观测数据似然函数的梯度信息都可以近似地计算. 因此, HMC 可以顺利实施.

令 $H(\boldsymbol{\theta},\boldsymbol{p}) = U(\boldsymbol{\theta}) + K(\boldsymbol{p})$ 代表 Hamilton 动力系统的能量方程,其中 $\boldsymbol{\theta}$ 是待估参数,$\boldsymbol{p}$ 是动量向量.算法 8 给出了 HABC 算法的流程.

算法 8 HABC 算法

1) 令 $t = 0$.从先验分布中产生初始参数 $\boldsymbol{\theta}^{(0)}$.

2) 令 $t = t + 1$.

3) 抽样初始动量参数 $\boldsymbol{p}^{(t-1)} \sim \pi_0(\boldsymbol{p})$.

4) 执行 L 步蛙跳算法(leap-frog),起点为 $[\boldsymbol{\theta}^{(t-1)},\boldsymbol{p}^{(t-1)}]$,步长为 δ_{LF}.最终得到 $\boldsymbol{\theta}^*,\boldsymbol{p}^*$.

5) 计算 $r = \min(1,\exp(-U(\boldsymbol{\theta}^*) + U(\boldsymbol{\theta}^{(t-1)}) - K(\boldsymbol{p}^*) + K(\boldsymbol{p}^{(t-1)})))$.

6) 令 u 代表标准均匀分布的一个随机数.如果 $u \leqslant r$,则令 $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$.否则,$\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^{(t-1)}$.

7) 返回步骤 2),直到算法的终止条件被满足.

步骤 4)中的蛙跳算法具体实施如下:

$$1)\ p_i(t+\delta_{\mathrm{LF}}/2)=p_i(t)-\frac{\delta_{\mathrm{LF}}}{2}\frac{\partial U}{\partial\theta_i(t)},$$

$$2)\ \theta_i(t+\delta_{\mathrm{LF}})=\theta_i(t)+\delta_{\mathrm{LF}}\frac{\partial K}{\partial\theta_i(t+\delta_{\mathrm{LF}}/2)},$$

$$3)\ p_i(t+\delta_{\mathrm{LF}})=p_i(t+\delta_{\mathrm{LF}}/2)-\frac{\delta_{\mathrm{LF}}}{2}\frac{\partial U}{\partial\theta_i(t+\delta_{\mathrm{LF}})},$$

其中 $\partial U/\partial\theta_i(t)$ 可以根据式(15)得到.

$$U(\boldsymbol{\theta})=-\ln(\pi_\delta(\boldsymbol{y}_n\mid\boldsymbol{\theta}))\approx-\ln\left((1/M)\sum_{m=1}^{M}\pi_\delta(\boldsymbol{y}_n\mid\boldsymbol{z}^{(m)})\right)\approx$$
$$\ln M+\ln\delta+(\boldsymbol{y}_n-\boldsymbol{z}^{(m^*)})^2/(2\delta^2),$$

其中 $\boldsymbol{z}^{(m^*)}$ 是距离观测数据最近的合成数据.

2.3.5 合成似然函数

合成似然函数(synthetic likelihood)方法是与 ABC 算法相近的算法之一.合成似然函数是由文献[49]提出.他的目标是对复杂的非线性模型做统计推断.由于观测数据具有高度的非线性和模型的复杂性,似然函数难以计算.他提出将观测数据 $\boldsymbol{y}_n$ 转化为一系列的统计量 $\boldsymbol{s}$.之后所有的统计推断都是基于统计量的空间进行.

在给定参数 $\boldsymbol{\theta}$ 时,通过生成大量合成数据 $\boldsymbol{z}_{1:N}$,计算相应的统计量 $\boldsymbol{s}_{\boldsymbol{\theta}}^{1:N}$.他假定统计量服从多元 Gauss 分布,即 $\boldsymbol{s}_{\boldsymbol{\theta}}\sim N(\boldsymbol{\mu}_{\boldsymbol{\theta}},\boldsymbol{\Sigma}_{\boldsymbol{\theta}})$.所以,借助 $\boldsymbol{s}_{\boldsymbol{\theta}}^{1:N}$ 估计 $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$.因此,对于特定的 $\boldsymbol{\theta}$,观测数据的似然函数是可以计算的.令 $\boldsymbol{\theta}_{\mathrm{obs}}$ 代表观测数据的相应统计量.$f(\boldsymbol{y}_n\mid\boldsymbol{\theta})\approx\phi(\boldsymbol{\theta}_{\mathrm{obs}}\mid\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}},\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\theta}})$,其中 $\phi(\cdot\mid\boldsymbol{\theta})$ 是多元正态分布的密度函数.因此,MCMC 可以用于从合成似然函数中获得后验样本.

合成似然函数与 ABC 的相同之处在于:它们都需要生成大量的合成数据.不同之处在于:合成似然函数假设合成数据的统计量服从特定的分布,而 ABC 则直接比较生成数据与观测数据之间的差异.显然合成似然函数并不是真正的似然函数.它使用一个易于计算的似然函数代替了难以计算的真正似然函数.本质上,ABC 方法不需要计算似然函数,但是合成似然函数方法是基于似然函数的推断方法.

3 ABC 方法在复杂数据处理中的应用

ABC 方法的上述前沿进展在一系列复杂数据处理问题中有着广泛的应用,并和对抗生成网络等前沿人工智能方法有着深刻的内在联系.本节将对相关案例和联系给予讨论.

3.1 ABC 方法与连续时间自回归

在信号处理、金融等领域经常会处理连续时间序列的数据类型.连续时间自回归模型(continuous-time autoregressive model,CAR)是处理这类数据的常用模型.经典 CAR 模型可以使用 Kalman 滤波(Kalman filtering)和粒子滤波(particle filter)做统计推断[67].然而,一旦 CAR 模型中的一些经典假设条件不再成立,其统计推断就会遇到很大的挑战.ABC 方法是解决这类挑战的一个有力工具[68].

通常,在状态空间形式下,P 阶 CAR 模型被如下描述:

$$\mathrm{d}\boldsymbol{X}(t)=\boldsymbol{A}\boldsymbol{X}(t)\mathrm{d}t+\tilde{\boldsymbol{e}}\mathrm{d}W(t),\tag{16}$$

其中,$\{W(t)\}$ 是标准 Brown 运动,$\boldsymbol{X}(t)=[X(t),X^{(1)}(t),\cdots,X^{(P-1)}(t)]^{\mathrm{T}}$($X^{(i)}(t)$ 表示 $X(t)$ 的第 i 阶导数),

$$\boldsymbol{A}=\begin{pmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ -a_P & -a_{P-1} & -a_{P-2} & \cdots & -a_1\end{pmatrix}, \tag{17}$$

$\tilde{\boldsymbol{e}}=[0,0,\cdots,\sigma_e]^{\mathrm{T}}$.模型的 Itô 积分解如下[69-70]:

$$\boldsymbol{X}(t)=\mathrm{e}^{\boldsymbol{A}t}\boldsymbol{X}(0)+\int_0^t \mathrm{e}^{\boldsymbol{A}(t-u)}\tilde{\boldsymbol{e}}\mathrm{d}W(u), \tag{18}$$

其中状态转移是 Gauss 密度函数:

$$f(\boldsymbol{X}(t)\mid \boldsymbol{X}(s))=\mathcal{N}(\mathrm{e}^{\boldsymbol{A}t}\boldsymbol{X}(s),C(t,s)),\qquad s<t, \tag{19}$$

$$C(t,s)=\mathrm{cov}(\boldsymbol{X}(t)\mid \boldsymbol{X}(s))=\int_s^t \mathrm{e}^{\boldsymbol{A}(t-u)}\tilde{\boldsymbol{e}}\tilde{\boldsymbol{e}}^{\mathrm{T}}\mathrm{e}^{\boldsymbol{A}^{\mathrm{T}}(t-u)}\mathrm{d}u. \tag{20}$$

连续时间过程可以放在如下离散状态空间:

$$X_{t_i}=\mathrm{e}^{\boldsymbol{A}(t_i-t_{i-1})}X_{t_{i-1}}+e_{t_i},\qquad i=1,2,\cdots,N, \tag{21}$$

其中系统噪声 $e_{t_i}\sim\mathcal{N}(0,\sigma_s^2)$,它的方差是

$$\sigma_s^2=\int_0^{t_{i+1}-t_i}\mathrm{e}^{\boldsymbol{A}\tau}\tilde{\boldsymbol{e}}\tilde{\boldsymbol{e}}^{\mathrm{T}}\mathrm{e}^{\boldsymbol{A}^{\mathrm{T}}\tau}\mathrm{d}\tau. \tag{22}$$

如果驱动噪声 $\{W(t)\}$ 是标准 Brown 运动,那么 Kalman 滤波方法可以精确求解似然函数[68].但是,在 $\{W(t)\}$ 不满足标准 Brown 运动的假设时,计算似然函数就尤其困难.我们考虑一阶 CAR 模型.令 $\boldsymbol{\theta}=(a_1,\sigma_e^2)$.观测数据为 $\boldsymbol{X}=(X_{t_1},X_{t_2},\cdots,X_{t_N})$.我们观察到,给定一组参数候选值 $\boldsymbol{\theta}^*$,可以根据式(21)计算 $z_{t_i}=X_{t_i}-\mathrm{e}^{\boldsymbol{A}^*(t_i-t_{i-1})}X_{t_{i-1}}$.如果 $\boldsymbol{\theta}^*$ 与真实值一致且 $t_1,t_2,\cdots,t_N$ 是等差数列,那么序列 z_{t_i} 是独立同分布的的样本集.因此,可以根据 z_{t_i} 是否为独立同分布来推测 $\boldsymbol{\theta}^*$ 与真实值的差异度.近年来,距离相关性(distance correlation)[71]是双样本检测的一个热门工具.我们拟使用距离相关性与 ABC 结合的方法推断 CAR 模型的参数.算法由文献[68]提出,具体的步骤由算法 9 展示.

算法 9 针对 CAR 模型的 ABCDC 算法

1) 令 0 时刻观测值为 $\boldsymbol{X}(0)=x$.

2) 从先验分布 $\pi(a)$ 中抽样一组候选值 a^*.

3) 如果 $P=1$,进入步骤 4).否则,使用数值方法获得 $\boldsymbol{X}(t)$ 的导数.

4) 计算 $z_{t_i}=X_{t_i}-\mathrm{e}^{\boldsymbol{A}(t_i-t_{i-1})}X_{t_{i-1}}$.

5) 将 $\{z\}$ 划分为两部分.令 $z^1=\{z_1,z_3,\cdots\}$ 以及 $z^2=\{z_2,z_4,\cdots\}$.

6) 计算距离相关性 $D\mathrm{cor}(z^1,z^2)$.

7) 如果 $D\mathrm{cor}(z^1,z^2)\leqslant\varepsilon$,接受 a^*.

8) 返回步骤 2),直至终止条件达成.

为了提高抽样效率,在实际应用中可以将上述算法与 MCMC 结合,形成 ABCDC-MCMC 算法.此处不再赘述算法细节,有兴趣的读者可以参考文献[68].相关方法的实际效果得到了数值模拟的验证.

Ornstein Uhlenbeck(OU)过程是连续时间序列模型的一类过程.不同的驱动过程定义不同的 OU 过程.图 5(a)给出了从如下 OU 过程中产生的一组模拟数据:

$$X_{t_i}=\mathrm{e}^{a(t_i-t_{i-1})}X_{t_{i-1}}+W_{t_i},\qquad i=1,2,\cdots,N, \tag{23}$$

其中，$W_{t_i}=\sigma_e\int_{t_{i-1}}^{t_i}\mathrm{e}^{a(t_i-u)}\mathrm{d}L(u)$，$L$ 代表驱动过程，参数的真值为 $a=-0.8$，$\sigma_e=1$.此处，令 L 表示 Brown 运动.

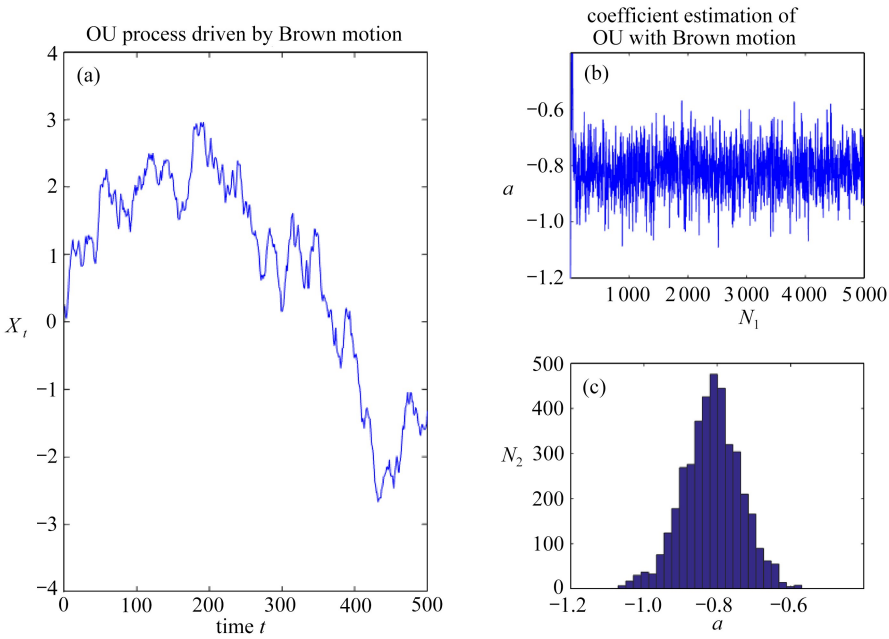

图 5 连续时间自回归模型的 Bayes 推断

Fig. 5 Bayesian inference of coutinuous-time autoregression model

图 5(b)、(c)展示了运用 ABCDC-MCMC 算法所得到关于参数 a 的近似后验分布，其中图 5(b)为参数 a 的 MCMC 样本，图 5(c)为参数 a 的直方图.相关结果表明 ABCDC-MCMC 方法能够有效推断参数 a.而 W_{t_i} 的方差为 $S^2(W_{t_i})=\sigma_e^2\int_0^{t_i-t_{i-1}}\mathrm{e}^{2a\tau}\mathrm{d}\tau$.因此，令 $\hat{a}$ 代表 a 的估计值，则

$$\hat{\sigma}_e=\sqrt{\frac{S^2(z)}{\int_0^{t_i-t_{i-1}}\mathrm{e}^{2\hat{a}\tau}\mathrm{d}\tau}}.$$

3.2 ABC 方法与人工智能

以生成对抗网络(GAN)和变分自动编码器(VAE)为代表的新型生成模型是近年来人工智能研究中的一大热点，其应用范围涵盖了图像、语音、文本等常见的复杂数据应用领域.给定一组希望被模仿的目标数据样本，这类方法力图输出一组生成数据样本，使得生成数据样本与目标数据样本尽可能相似而不能被区分.不同于在应用数学和统计学中广泛使用的基于领域知识进行建模的传统生成模型，这些新型生成模型以数据驱动的方式，通过人工模拟的手段，对数据生成过程进行探索和刻画.作为对生成模型进行统计推断的一般性方法，ABC 方法原则上可以与 GAN 和 VAE 深度结合，在人工智能研究中引入不确定推断的理念和方法，并运用前沿人工智能技术推动不确定性推断方法的不断发展.

一般而言，GAN 由一个生成网络(generative network)和一个判别网络(discriminative network)构成，其中生成网络用于产生生成数据样本，判别网络用于区分生成数据样本和目标数据样本.GAN 方法的核心在于：通过生成网络和判别网络之间迭代进行的对抗训练来不断提高两者的能力，并最终使得生成网络输出的生成数据样本达到可以乱真的程度.所谓的对抗训练

一般由两个步骤组成:首先,训练生成网络,使得其生成的样本可以骗过当前的判别网络,被判别为和目标数据样本属于同一类别;然后,训练并更新判别网络,使其能够以更大的概率正确区分生成样本和目标样本.容易看到,通过迭代式的对抗训练,生成网络和判别网络的能力均可不断提升,并最终到达某种平衡.令 $\mathcal{G}$ 表示所有可容许的生成网络所构成的空间,$\mathcal{D}$ 表示所有可容许的判别网络所构成的空间,上述通过迭代对抗训练来改进直观生成及判别网络的过程可以通过求解如下优化问题来实现[9]:

$$\min_{G\in\mathcal{G}}\max_{D\in\mathcal{G}} \boldsymbol{V}(G,D),$$

其中目标函数:

$$\boldsymbol{V}(G,D) \triangleq E_{\boldsymbol{y}\sim p_{\text{data}}(\boldsymbol{y})}[\ln(D(\boldsymbol{y}))] + E_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\ln(1-D(G(\boldsymbol{z})))], \tag{24}$$

其中 $p_{\text{data}}(\boldsymbol{y})$ 代表观测数据的真实分布,$p_z(\boldsymbol{z})$ 代表生成网络输入的分布.有进一步的理论研究发现:上述优化问题等效于最小化由生成网络产生的生成数据样本的概率分布 F_G 与目标数据所对应的概率分布 F_T 之间的 KL 散度[9].考虑到当 F_G 与 F_T 的支持空间重叠度较小或无重叠的时候,这种基于 KL 散度的优化策略常常无法提供可引导优化算法的有效信号(即所谓"梯度消失"现象),从而造成对抗训练的失败,人们开始尝试使用一系列基于双样本检测(two sample test)的统计量(包括 Wasserstein 距离[16]、最大平均差异[72]、能量距离和 χ^2 距离[19]等)来替代 KL 散度作为度量生成数据样本与目标数据样本相似程度的度量.相关改进产生了一系列效果非常好的新型 GAN 算法,在一系列应用问题中取得了显著的效果.

作为生成模型的另一种典型代表,VAE[10]由一个编码器(encoder)和一个解码器(decoder)构成.其中,编码器将输入的目标数据样本转换为近似服从 d-维标准正态分布 $\boldsymbol{N}_d$ 的一组隐变量 $\{H_i\}_{i=1}^n$,并用这些隐变量及其分布 $\boldsymbol{N}_d$ 表示目标数据样本的核心特征;然后,通过从分布 $\boldsymbol{N}_d$ 中抽取新的隐变量 H^* 并透过解码器对 H^* 进行解码即可产生一个新的生成数据样本 $\boldsymbol{y}^*$.事实上,变分自动编码器是"变分法"(variational Bayesian)的一项重要应用.变分法是针对大规模模型和海量数据的 Bayes 推断算法的改进.特别是随机变分法(stochastic variational inference)的提出[73],它放弃了推导解析变分公式的复杂过程,改用随机梯度下降(stochastic gradient descent, SGD)算法求解变分分布,使得这类变分法特别适合解决大规模 Bayes 模型.

受 GAN 和 ABC 算法的影响,近期 VAE 模型也在尝试用统计量的度量做似然函数的近似,如文献[74]用 MMD 等统计量替代 VAE 中的分布解析表达式;如文献[75]提出对抗自动编码器,将 VAE 的变分推断与 GAN 中判别网络的统计量度量结合在一起.此外,变分法的发展,也给 ABC 的 Bayes 推断过程提供了新的思路.文献[76]提出了变分 ABC,用变分推断替代传统 ABC 中基于 Monte-Carlo 抽样的后验分布估计.也可以说是用统计量度量替代传统变分推断中的似然函数,为变分推断和 ABC 算法的融合做出了典范.

其实从某种意义上来说,GAN 及其变种模型属于泛化的 ABC 算法.它们使用深度神经网络作为样本数据生成器,然后使用深度神经网络作为判别器.此处判别网络先将数据变换到一个较低维度的特征向量,然后用生成样本的特征向量与目标样本的特征向量之间的距离作为度量,并以此度量来设计损失函数.与 ABC 不同的之处有二:第一,GAN 模型泛化了 ABC 方法的数据生成方式,ABC 方法的数据生成方式是已知的,GAN 模型需要学习数据的生成方式.第二,经典的 ABC 模型需要使用者提供数据的统计量,GAN 方法是通过判别网络学习数据的特征向量.可以预见 ABC 领域的理论发展,特别是构建高维数据的统计量、统计量的度量和判别等方面的进步,将进一步促进 GAN 网络等生成模型的进步;同时 GAN 网络等新兴生成模型的

进步,也给 ABC 在大规模、复杂数据领域的应用提供了新的思路.

以上关于 GAN 和 VAE 的模型大多数只注重得到模型的“最优解”.在众多场景下,如医疗[77]、自动驾驶(对结果判断可靠性要求比较高)“最优解”往往并不足够.然而,不确定性推断可以帮助获得关于参数的更全面信息.ABC 算法可以对模型的不确定性进行量化分析,使模型给出更加安全可靠的推断.可以预见 ABC 中用样本统计量替代解析似然函数的哲学理念,融合了深度神经网络的复杂而灵活的统计模型和基于变分法的大规模 Bayes 推断算法三者的相互融合,将成为人工智能领域研究的热点,也会使得 ABC 类型的算法在机器学习领域获取更大的应用空间.另外,ABC 领域的理论发展,特别是构建高维数据的统计量、统计量的度量和判别等方面的进步,将进一步促进 GAN 网络等生成模型的进步;同时 GAN 网络等新兴生成模型的进步,也给 ABC 在大规模、复杂数据领域的应用提供了新的思路.

4 结 论

生成模型能够有效处理复杂数据的模型和算法以从数据中获取有用的信息和知识.它已经成为机器学习和统计学习中的重要研究领域.虽然生成模型在大数据时代拥有巨大的优势,但是目前的生成模型并不重视不确定性分析.近似 Bayes 计算既拥有生成模型的优势,又能实现不确定性分析.因此,本文期望能够将近似 Bayes 计算的优势与深度学习技术结合,发展出更加完善的生成模型算法,以期它能够在大数据时代发挥作用.首先,本文回顾了 ABC 的产生和发展历程.其次,对于 ABC 与其他技术的关系和扩展也做了比较详细的描述.ABC 的发展主要集中在三个方面:构建统计量、选择距离度量和选择阈值.这三个方面制约了 ABC 的应用也在推动 ABC 的发展.然后,我们介绍了 ABC 技术在大数据时代的应用潜力.

总体而言,ABC 框架已经成为统计计算和统计学习领域的一个重要分支.在统计领域,它可以处理复杂模型,如生态模型、生物模型和传染病模型等.在当今的大数据时代,模型变得更加复杂,基于似然函数的方法遇到很大挑战.ABC 与其他大数据技术逐渐结合,它们在处理复杂数据和模型时拥有巨大优势,相信 ABC 会在大数据时代发挥更加重要的作用.

致谢 本文作者衷心感谢北京智源人工智能研究院对本文的支持.

参考文献(References):

[1] STRANG G. *Introduction to Applied Mathematics*[M]. Wellesley, MA: Wellesley-Cambridge Press, 1986.

[2] 谷超豪, 李大潜, 陈恕行. 数学物理方法[M]. 上海: 上海科学技术出版社, 2002.(GU Chaohao, LI Daqian, CHEN Shuxing. *Mathematical and Physical Methods*[M]. Shanghai: Shanghai Scientific & Technical Publishers, 2002.(in Chinese))

[3] BOARD A. *Stochastic Modelling and Applied Probability*[M]. Springer, 2005.

[4] COURANT R, HILBERT D. *Methods of Mathematical Physics: Partial Differential Equations*[M]. John Wiley & Sons, 2008.

[5] ARNOL'D V I. *Mathematical Methods of Classical Mechanics*[M]. Springer Science & Business Media, 2013.

[6] NETER J, KUTNER M H, NACHTSHEIM C J, et al. *Applied Linear Statistical Models*[M]. Chicago, 1996.

[7] CONGDON P. *Bayesian Statistical Modelling*[M]. John Wiley & Sons, 2007.

[8] FAHRMEIR L, TUTZ G. *Multivariate Statistical Modelling Based on Generalized Linearmod-*

els[M]. Springer Science & Business Media, 2013.

[9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//*Advances in Neural Information Processing Systems 27* (*NIPS 2014*). Montreal, Canada, 2014.

[10] KINGMA D P, WELLING M. Auto-encoding variational Bayes[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1511.06434.pdf.

[11] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1312.6114.pdf.

[12] CHEN X, DUAN Y, HOUTHOOFT R, et al. Infogan: interpretable representation learning by information maximizing generative adversarial nets[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1606.03657.pdf.

[13] ZHANG Han, XU Tao, LI Hongsheng, et al. Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1612.03242v1.pdf.

[14] MAO X D, LI Q, XIE H R, et al. Least squares generative adversarial networks[C]//*2017 IEEE International Conference on Computer Vision* (*ICCV*). 2017.

[15] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//*2017 IEEE Conference on Computer Vision and Pattern Recognition*(*CVPR*). 2017: 1125-1134.

[16] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein GAN[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1701.07875.pdf.

[17] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//*2017 IEEE International Conference on Computer Vision* (*ICCV*). Venice, Italy, 2017.

[18] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]//*2019 IEEE Conference on Computer Vision and Pattern Recognition*(*CVPR*). 2019: 4401-4410.

[19] TAO C Y, CHEN L Q, HENAO R, et al. Chi-square generative adversarial network[C]//*International Conference on Machine Learning*. 2018: 4894-4903.

[20] SØNDERBY C K, RAIKO T, MAALØE L, et al. Ladder variational autoencoders[C]//*Advances in Neural Information Processing Systems 29* (*NIPS 2016*). 2016: 3738-3746.

[21] HIGGINS I, MATTHEY L, GLOROT X, et al. Early visual concept learning with unsupervised deep learning[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1606.05579.pdf.

[22] RUBIN D B. Bayesianly justifiable and relevant frequency calculations for the applies statistician[J]. *The Annals of Statistics*, 1984, **12**(4): 1151-1172.

[23] PRITCHARD J K, SEIELSTAD M T, PEREZ-LEZAUN A, et al. Population growth of human Y chromosomes: a study of Y chromosome microsatellites[J]. *Molecular Biology & Evolution*, 1999, **16**(12): 1791-1798.

[24] WILKINSON R D,TAVARÉB S. Estimating primate divergence times by using conditioned birth-and-death processes[J]. *Theoretical Population Biology*, 2009, **75**(4): 278-285.

[25] PETERS G W, SISSON S A, FAN Y. Likelihood-free Bayesian inference for Alpha-stable models[J]. *Computational Statistics & Data Analysis*, 2012, **56**(11): 3743-3756.

[26] NOTT D J, FAN Y, MARSHALL L, et al. Approximate Bayesian computation and Bayes' linear analysis: toward high-dimensional ABC[J]. *Journal of Computational and Graphical Sta-*

tistics, 2014, **23**(1): 65-86.

[27] RATMANN O, ANDRIEU C, WIUF C,et al. Model criticism based on likelihood-free inference, with an application to protein network evolution[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**(26): 10576-10581.

[28] KULKARNI T, YILDIRIM I, KOHLI P,et al. Deep generative vision as approximate Bayesian computation[C]//*NIPS 2014 ABC Workshop*. 2014.

[29] SHEEHAN S, SONG Y S. Deep learning for population genetic inference[J]. *PLoS Computational Biology*, 2016, **12**(3): e1004845. DOI: 10.1371/journal.pcbi.1004845.

[30] GAL Y, GHAHRAMANI Z B. Dropout as a Bayesian approximation: representing model uncertainty in deep learning[C]//*Proceedings of the 33rd International Conference on Machine Learning*. New York, USA, 2016: 1050-1059.

[31] FELIP J, AHUJA N, GOMEZ-GUTIERREZ D, et al. Real-time approximate Bayesian computation for scene understanding[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1905.13307.pdf.

[32] GHOSH J K, RAMAMOORTHI R V. *Bayesian Nonparametrics*[M]. New York: Springer, 2003.

[33] ROBERT C. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*[M]. Springer Science, 2007.

[34] BERNARDO J M, SMITH A F M. *Bayesian Theory*[M]. John Wiley & Sons, 2009.

[35] GELMAN A, CARLIN J B, STERN H S, et al. *Bayesian Data Analysis*[M]. 3rd ed. Chapman and Hall/CRC, 2013.

[36] LIU J S. *Monte Carlo Strategies in Scientific Computing*[M]. New York: Springer, 2001.

[37] BROOKS S, GELMAN A, JONES G, et al. *Handbook of Markov Chain Monte Carlo*[M]. New York: CRC press, 2011.

[38] CHEN M H, SHAO Q M, IBRAHIM J G. *Monte Carlo Methods in Bayesian Computation*[M]. Springer Science & Business Media, 2012.

[39] GIUDICI P, GIVENS G H, MALLICK B K. *Wiley Series in Computational Statistics*[M]. Wiley Online Library, 2013.

[40] MARJORAM P, MOLITOR J, PLAGNOL V, et al. Markov chain Monte Carlo without likelihoods[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2003, **100**(26): 15324-15328.

[41] SISSON S A, FAN Y, TANAKA M M. Sequential monte carlo without likelihoods[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2007, **104**(6): 1760-1765.

[42] FEARNHEAD P, PRANGLE D. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation[J]. *Journal of the Royal Statistical Society: Series B* (*Statistical Methodology*), 2012, **74**(3): 419-474.

[43] DELMORAL P, DOUCET A, JASRA A. An adaptive sequential Monte Carlo method for approximate Bayesian computation[J]. *Statistics and Computing*, 2012, **22**(5): 1009-1020.

[44] MENGERSEN K L, PUDLO P, ROBERT C P. Bayesian computation via empirical likelihood [J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2013, **110**(4): 1321-1326.

[45] SISSON S A, FAN Y, BEAUMONT M. *Handbook of Approximate Bayesian Computation*[M]. Chapman and Hall/CRC, 2018.

[46] FRAZIER D T, MARTIN G M, ROBERT C P. On consistency of approximate Bayesian computation[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1508.05178.pdf.

[47] HEIN J, SCHIERUP M, WIUF C. *Gene Genealogies, Variation and Evolution: a Primer in Coalescent Theory*[M]. Oxford: Oxford University Press, 2004.

[48] TANAKA M M, FRANCIS A R, LUCIANI F, et al. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data[J]. *Genetics*, 2006, **173**(3): 1511-1520.

[49] WOOD S N. Statistical inference for noisy nonlinear ecological dynamic systems[J]. *Nature*, 2010, **466**(7310): 1102-1104.

[50] ZHU W C, FAN Y. A novel approach for Markov random field with intractable normalizing constant on large lattices[J]. *Journal of Computational and Graphical Statistics*, 2018, **27**(1): 59-70.

[51] PRANGLE D. *Summary Statistics*[M]. Chapman and Hall/CRC, 2018.

[52] JOYCE P, MARJORAM P. Approximately sufficient statistics and Bayesian computation[J]. *Statistical Applications in Genetics and Molecular Biology*, 2008, **7**(1). DOI: 10.2202/1544-6115.1389.

[53] HASTIE T, TIBSHIRANI R, FRIEDMAN J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*[M]. 2nd ed. New York: Springer, 2009.

[54] MARDIA K V, KENT J T, BIBBY J M. Multivariate analysis[M]//*Probability and Mathematical Statistics*. New York: Academic Press, 1979.

[55] JIANG B, WU T Y, ZHENG C, et al. Learning summary statistic for approximate Bayesian computation via deep neural network[J]. *Statistica Sinica*, 2017, **27**(4): 1595-1618.

[56] CREEL M. Neural nets for indirect inference[J]. *Econometrics and Statistics*, 2017, **2**: 36-49.

[57] RAYNAL L, MARIN J M, PUDLO P, et al. ABC random forests for Bayesian parameter inference[J]. *Bioinformatics*, 2018, **35**(10): 1720-1728.

[58] BEAUMONT M A. Approximate Bayesian computation[J]. *Annual Review of Statistics and Its Application*, 2019, **6**(1): 379-403.

[59] HARRISON J U, BAKER R E. An automatic adaptive method to combine summary statistics in approximate Bayesian computation[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1703.02341.pdf.

[60] BERNTON E, JACOB P E, GERBER M, et al. Approximate Bayesian computation with the wasserstein distance[J]. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2019, **81**(2): 235-269.

[61] PARK M, JITKRITTUM W, SEJDINOVIC D. K2-ABC: approximate Bayesian computation with kernel embeddings[C]//*Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Cadiz, Spain, 2016.

[62] SISSON S A, FAN Y, TANAKA M M. Sequential monte carlo without likelihoods: errata[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2009, **106**(39): 1760-1765.

[63] BEAUMONT M A, CORNUET J M, MARIN J M, et al. Adaptive approximate Bayesian computation[J]. *Biometrika*, 2009, **96**(4): 983-990.

[64] TONI T, WELCH D, STRELKOWA N, et al. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems[J]. *Journal of the Royal Society Interface*, 2009, **6**(31): 187-202.

[65] DUANE S, KENNEDY A D, PENDLETON B J, et al. Hybrid Monte Carlo[J]. *Physics letters B*, 1987, **195**(2): 216-222.

[66] MEEDS E, LEENDERS R, WELLING M. Hamiltonian ABC[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1503.01916.pdf.

[67] GIANNOPOULOS P, GODSILL S J. Estimation of car processes observed in noise using Bayesian inference[C]//*2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Salt Lake City, UT, USA, 2001.

[68] JI C L, YANG L G, ZHU W C, et al. On Bayesian inference for continuous-time autoregressive models without likelihood[C]//*2018 21st International Conference on Information Fusion (FUSION)*. 2018: 2137-2142.

[69] HARVEY A C. *Forecasting, Structural Time Series Models and the Kalman Filter*[M]. Cambridge: Cambridge University Press, 1991.

[70] JONES R H. Fitting a continuous time autoregression to discrete data[C]//*Proceedings of the Second Applied Time Series Symposium Held*. Tulsa, Oklahoma, 1980.

[71] SZÉKELY G J, RIZZO M L, BAKIROV N K. Measuring and testing dependence by correlation of distances[J]. *The Annals of Statistics*, 2007, **35**(6): 2769-2794.

[72] LI C L, CHANG W C, CHENG Y, et al. MMD GAN: towards deeper understanding of moment matching network[C]//*Advances in Neural Information Processing Systems 30 (NIPS 2017)*. 2017.

[73] HOFFMAN M D, BLEI D M, WANG C, et al. Stochastic variational inference[J]. *Journal of Machine Learning Research*, 2013, **14**(1): 1303-1347.

[74] ZHAO S J, SONG J M, ERMON S. Infovae: information maximizing variational autoencoders[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1706.02262.pdf.

[75] MAKHZANI A, SHLENS J, JAITLY N, et al. Adversarial autoencoders[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1511.05644.pdf.

[76] MORENO A, ADEL T, MEEDS E, et al. Automatic variational ABC[R/OL]. [2019-08-26]. https://arxiv.org/pdf/1606.08549.pdf.

[77] 王小娥, 蔺小林, 李建全. 一类具有脉冲免疫治疗的 HIV-1 感染模型的动力学分析[J]. 应用数学和力学, 2019, **40**(7): 728-740.(WANG Xiaoe, LIN Xiaolin, LI Jianquan. Dynamic analysis of a class of HIV-1 infection models with pulsed immunotherapy[J]. *Applied Mathematics and Mechanics*, 2019, **40**(7): 728-740.(in Chinese))

Recent Progress of Approximate Bayesian Computation and Its Applications

ZHU Wanchuang[1], JI Chunlin[2], DENG Ke[1]

(1. *Center for Statistical Science*, *Department of Industrial Engineering*, *Tsinghua University*, *Beijing* 100084, *P.R.China*;
2. *Kuang-Chi Institute of Advanced Technology*, *Shenzhen*, *Guangdong* 518000, *P.R.China*)

Abstract: In the era of big data and artificial intelligence, it is a common challenge for applied mathematics, statistics and computer science to extract valuable information and knowledge from complex data and models. Generative models are a class of powerful models which can potentially handle the above difficulty. From a macro point of view, the differential equations and dynamic systems in applied mathematics, the probability distribution in statistical models, and the typical generative models (generative adversarial networks and variational auto-encoders) in computer science could be considered as generalized generative models. Along with larger and larger-size data, the structure of data becomes more and more complicated simultaneously. Therefore, more powerful generative models are essential to process real problems. It is a challenge to describe mathematical structures of these generative models. It poses a natural question of how to analyze such generative models without analytic forms (or hard to obtain their analytic forms). Originated from the Bayesian inference, the approximate Bayesian computation, as a likelihood-free technique, plays an important role in processing complex statistical models and generative models. Based on the classic approximate Bayesian computation, the development and recent advance of approximate Bayesian computation were systematically reviewed. Finally, the application of the approximate Bayesian computation to complex data and the deep connection between the approximate Bayesian computation and cutting-edge artificial intelligence methods were discussed.

Key words: approximate Bayesian computation; generative model; deep learning; uncertainty inference

Foundation item: The National Natural Science Foundation of China(11771242)

引用本文/Cite this paper:

朱万闯, 季春霖, 邓柯. 近似 Bayes 计算前沿研究进展及应用[J]. 应用数学和力学, 2019, **40**(11): 1179-1203.

ZHU Wanchuang, JI Chunlin, DENG Ke. Recent progress of approximate Bayesian computation and its applications[J]. *Applied Mathematics and Mechanics*, 2019, **40**(11): 1179-1203.